REPORT ON
THE DATA QUALITY INDICATORS WORKGROUP

In early 1999, Mark Day, Acting Director of the Office of Information Resources Management (now in The Office Technology Operations and Planning of The Office of Environmental Information, OEI), invited National Program Manager's Quality Assurance Representatives and Senior Information Resources Management Officers, as well as the Regional Information Resources Management Branch Chiefs to identify representatives for an Environmental Data Registry workgroup developed jointly by the Office of Information Resources Management (OIRM) and the Quality Assurance Division (QAD) of ORD (now the Quality Staff (QS) of the OEI). The invitation was an attempt to involve all stakeholders up-front to ensure broad acceptance of the workgroup's efforts. The workgroup was tasked to develop standard definitions and formats for data elements for quality assurance terms used across the Agency to describe data quality such as "precision." Once developed through consensus, these definitions and formats would be stored in the Web-based Environmental Data Registry (EDR), the Agency's single comprehensive source of information about environmental data, where they will be available for re-use in regulations, quality assurance plans, and legislation, as well as to enhance public understanding of the meaning of EPA's data.

Linda Kirkland of QAD and Tom Maloney of OIRM had spoken about the workgroup with Agency staff known to be interested in data quality issues, and several have volunteered to serve. Later Regional, State and other representatives were added from responses to Mark Day's memorandum:

Data Quality Indicators Work Group

Linda Kirkland QS/OEI (Chair)	Tom Maloney/Brand Niemann OEI
Dallas Wright OPPTS	Duane Geuder OSWER
Doris Maxwell OAR/OAQPS	Wendy Blake-Coleman OW
ORD Lora Johnson	Joan Fisk QAEM (data transfer standard)
Moira Lataille R1	Patricia Sheridan/Bob Runyon R2
Cheng-wen Tsai R5	Timothy Dawson R6
Doug Brune R7	Bill Monson R8
Vance Fong/Roseanne Sakamoto R9	Don Matheny R10
Martin Topper OECA	David Eng/ Tony Jover OSWER
Barbara Hughes NEIC	James Rothwell, OIG

Danile Chang Hawaii	Ken Carlson AZ
Frances Haertel R6/OGWDW	Chuck Job OGWDW
John Warren, QS/OEI	Jeanne Campbell OW
Brad Parsons CA	

Linda Kirkland agreed to serve as chairperson of the workgroup, and Tom provided support and coordination until December when Brand Niemann took over. The initial meeting developed the scope of the effort emphasizing the difference between data elements for data set metadata and the existing glossary of Quality Assurance terms. Subsequently, the following charter was developed to provide the mission statement, and a timetable for the effort.

Data Quality Indicators Workgroup Charter

Problem:

Enhanced public access and subsequent use/reuse of environmental measurement data has underscored the need to provide adequate information to characterize the quality of data within Agency databases. Data quality indicators are necessary to aid in determining if the use/reuse of a data set is appropriate. However, terms, indicators, and statistical formulae used to describe data quality are inconsistently defined and applied across the Agency. A consensus needs to be developed to promote data sharing between Agency programs and multi-program uses. Other factors influencing standardization are the requirements of EPA Order 5360.1, Change 1, Policy and Program Requirements for the Mandatory Agency-Wide Quality System, the Government Performance and Results Act, the Federal Geographic Data Committee, and the need for supporting common electronic data transfer.

Goals

The **short term goals** of the workgroup are to:

-- determine potential data quality information (elements) that should travel with a data set to characterize the quality of that data set, thereby ensuring that all users of secondary data will be able to determine whether those data meet their needs.
-- develop consensus definitions for commonly used data quality indicators (e.g., precision, accuracy, bias).
-- develop consensus statistical definitions/formulae used to measure data quality (e.g., relative percent difference, standard deviation).

The **long term goals** of the workgroup are to:

-- make consensus definitions publicly available on the Environmental Data Registry (EDR) Web site (www.epa.gov/edr).
-- promote the standardization and use of consensus definitions in the promulgation of federal environmental regulations and legislation.
-- promote the use of consensus data quality indicators and related terms by other federal agencies, state, tribal, local and industrial partners for use in developing required Quality Management Plans and Quality Assurance Project Plans.

The workgroup met every three or four weeks during 1999 to develop the definitions and formats for the EDR spreadsheet. following their proposed schedule:

Determine the minimum set of data quality indicators: September 1999
Define data elements which comprise the data quality indicators and their value domains incorporate into the EDR format December 1999
Data elements registered by OEI in Environmental Data Registry: March 2000
OEI & Workgroup publicize availability of elements: March 2000

The workgroup recognizes that to develop business rules for application of these data elements to data set objects, a pilot would be needed to demonstrate their use with the Environmental Information Management System, the Agency's metadata warehouse described below to meet Agency Quality system requirements. This will be undertaken after the data elements have completed review through the EDR website.

Quality System Requirements for Assessment of Database Data Sets to Support Use

The EPA Order 5360.1 Change 1 requirements cover "the use of environmental data collected for other purposes, or from other sources (also termed "secondary data"), including literature, industry surveys, compilations from computerized databases and information systems, results from computerized or mathematical models of environmental processes and conditions". Quality System requirements include "assessment of existing data, when used to support Agency decisions or other secondary purposes, to verify that they are of sufficient quantity and adequate quality for their intended use. Systematic planning is required "to develop acceptance or performance criteria for all work covered by the order". As described in Section 3.3.8 of the EPA Quality Manual for Environmental Programs, it is necessary in planning a project to identify the type of data needed and how the data will be used to support the project objectives. This is further refined in determining the quantity of data needed and the specifications of performance criteria for measuring quality. These criteria are used when planning how the acquired data will be analyzed and assessed against its intended use and quality performance criteria. Therefore, information on the quality of data sets stored in Agency data bases is necessary for it to be used by those who did not collect it. The Data Quality Indicators workgroup has identified and defined critical data quality indicators for data sets which can be provided as metadata for data set objects to assist users in locating existing data that may be acceptable for a defined project use. Qualitative descriptive data elements were developed for data set screening such as text about what the data represents in the environment and what is known about data set accuracy. Also data quality indicators were developed to summarize the quantitative quality control sample results for the data points in data sets so that their quality can be screened for characteristics such as completeness (is the quantity of data needed available?) and analytical accuracy (how close were the analytical results to a spike in the reference sample? how precise were those recoveries?).

Work Group Focus on a Data Element Group of Data Quality Indicators for Data Sets

The Data Set Data Quality Indicators Group is proposed to provide data element definitions for metadata describing the quality of data sets in the Environmental Information Management System (EIMS) and other related data bases. The EIMS is designed to facilitate environmental assessment activities conducted by the U.S. Environmental Protection Agency through the efficient capture, storage, management, and distribution of metadata and data. The metadata component of the system helps users identify, find, and evaluate environmental

resources collected, developed, and used by the Agency and its regional and state partners. The robust definition of metadata used within EIMS provides the functionality of describing multiple types of environmental resources, including: projects, data sets, databases, models, documents, and multimedia materials of potential use in environmental assessments. These metadata are provided to users through an interactive web interface that is dynamically linked to the relational database that is the core of the EIMS system. Additionally, a web interface is provided enabling USEPA staff, stakeholders, and state partners to contribute to the directory of environmental resources and related descriptive data elements stored within EIMS. Metadata are directly linked to the data component of EIMS. In this context, data may reside within the relational database, or be broadly distributed across the Internet in the form of individual data files. This integration between data and metadata provides the ability to link environmental data with the analytical, statistical, and modeling tools used to conduct environmental assessments. The system currently being used by the USEPA's Office of Water, Office of Air and Radiation, The Office of Research and Development, Region 10, and the Agency's State One-Stop Program.

The current architectural model for EIMS details a central relational database for the storage and management of the descriptive information about the types of resources used in the assessment process including: data sets, databases, models, documents, multimedia products, and web sites (Frithsen, 1999). The captured descriptive information is called metadata (literally, data about data), and, within EIMS, the assessment resources are called metadata objects. Data, documents, and models described by these metadata are distributed and may reside in or alongside of the relational database, on any server connected to the USEPA's Intranet, or on any Internet server, or physically in a library or investigator's desk drawer. At least at the highest level, this architecture is consistent with the model described for the Long-Term Ecological Research (LTER) Program, sponsored by the National Science Foundation; however, the vision for EIMS is carried a step further to reflect the need for the integration of metadata, data, and various assessment tools. This vision is described in the ORD Information Management Implementation Coordination Plan (Shepanek et al. 1999) and includes the option in the future of a distributed architecture for both data and metadata.

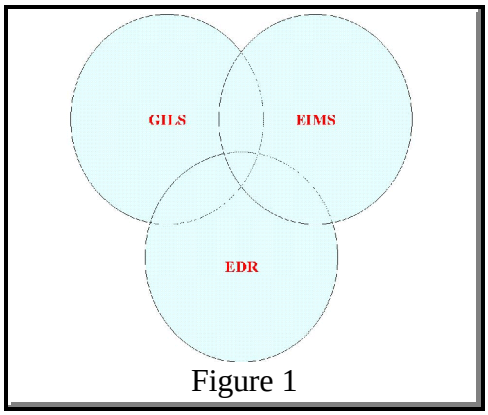

Figure 1

One of the primary functions of EIMS is to assist with the capture, management, and dissemination of metadata. Initial requirements and specifications for EIMS metadata were based upon existing guidelines developed for the NASA Master Directory, Directory Interchange Format (DIF), and the spatial metadata content standard developed by the Federal Geographic Data Committee (FGDC) (FGDC 1997). However, these requirements and specifications have been extended to reflect the needs of specific EIMS partners, and the vision for an integrated information management environment for ORD as outlined in the ORD Information Management Implementation Coordination Plan (Shepanek et al. 1999). Further EIMS requirements and specifications reflect the ultimate goal of being a central component of an integrated vision for metadata management within the Agency. This vision includes integration of the Federal Global Information Locator System (GILS), the USEPA's Environmental Data Registry (EDR), and EIMS (Figure 1).

Related Guidance and References:
Available at http://es.epa.gov/ncerqa/qa/qa_docs.html

EPA Order 5360.1 CHG 1 (1998): Policy and Program Requirements for the Mandatory Agency-wide Quality System

EPA Order 5360.1 CHG 1 (1998) defines the quality requirements for EPA organizations that produce environmental data. This Order replaces EPA Order 5360.1 in its entirety.

Current Version: Order 5360.1 CHG 1 (1998)
Contact: Quality Staff, (202) 564-6830

EPA Order 5360 (1998), EPA Quality Manual for Environmental Programs

The EPA Quality Manual defines program requirements for EPA organizations in implementing the mandatory
Quality System defined in EPA Order 5360.1 CHG 1. Equivalent specifications are defined in Requirements
Documents for organizations receiving financial assistance from EPA through extramural agreements.

Current Version: Order 5360 (1998)
Contact: Quality Staff, (202) 564-6830

EPA QA/G-4: Guidance for the Data Quality Objectives Process

QA/G-4 provides guidance on developing Data Quality Objectives using a systematic planning process based on a graded approach. The guidance presents a step-by-step description of the DQO process.

Current Version: Peer Review Version of Update to 9/94 Version
OR Final - EPA/600/R-96/055, September 1994
Contact: Quality Staff, (202) 564-6830

EPA QA/G-5: Guidance on Quality Assurance Project Plans

QA/G-5 provides guidance on developing Quality Assurance Project Plans (QAPPs) that will meet EPA requirements.

Current Version: Final - EPA/600/R-98/018, February 1998
Contact: Quality Staff, (202) 564-6830

EPA QA/G-9: Guidance for Data Quality Assessment: Practical Methods for Data Analysis

QA/G-9 provides guidance for planning and implementing assessments of environmental data. This document describes a method to perform a statistically-based, quantitative evaluation of the extent to which a data set satisfies the user's needs. The QA97 Version is now available.

Current Version: QA97 Version, EPA/600/R-96/084, January 1998 QA97 Update - contains 18 pages to update QA96 Version
Contact: Quality Staff, (202) 564-6830

EPA QA/G-4HW: Guidance for the Data Quality Objectives Process for Hazardous Waste Sites

QA/G-4HW provides guidance on applying the Data Quality Objectives (DQO) process to hazardous waste site investigations. The guidance presents a step-by-step description of the DQO process and its application to sampling designs for environmental remediation and waste management activities. Peer review is completed and this document is being finalized.

Current Version: Final - EPA/600/R-00/007, January 2000
Contact: Quality Staff, (202) 564-6830.

EPA QA/G-5i: Guidance on Data Quality Indicators

This guidance is under development and expect to be completed in 2000.

Frithsen, Jeffery B., 1999. Requirements for the Metadata Object - Models, EIMS Discussion Paper.

Shepanek, Robert F., G. J. Foley, G. B. Collins, L. L. Kirkland, J. H. Novak, J. B. Frithsen, M. Waters and collaborators. 1999. Scientific Information Management Implementation Coordination Plan 1999-2003. February 10, 1999. Office of Research and Development, U. S. Environmental Protection Agency, Washington, DC.